Extending Czech Aspect-Based Sentiment Analysis with Opinion Terms: Dataset and LLM Benchmarks

Jakub Šmíd[*,†], **Pavel Přibáň**[*], **Pavel Král**[*,†]
[*]Department of Computer Science and Engineering
[†]NTIS – New Technologies for the Information Society
University of West Bohemia in Pilsen, Faculty of Applied Sciences
Univerzitní 2732/8, 301 00 Pilsen, Czech Republic
{jaksmid, pribanp, pkral}@kiv.zcu.cz
https://nlp.kiv.zcu.cz

Abstract

This paper introduces a novel Czech dataset in the restaurant domain for aspect-based sentiment analysis (ABSA), enriched with annotations of opinion terms. The dataset supports three distinct ABSA tasks involving opinion terms, accommodating varying levels of complexity. Leveraging this dataset, we conduct extensive experiments using modern Transformer-based models, including large language models (LLMs), in monolingual, cross-lingual, and multilingual settings. To address cross-lingual challenges, we propose a translation and label alignment methodology leveraging LLMs, which yields consistent improvements. Our results highlight the strengths and limitations of state-of-the-art models, especially when handling the linguistic intricacies of low-resource languages like Czech. A detailed error analysis reveals key challenges, including the detection of subtle opinion terms and nuanced sentiment expressions. The dataset establishes a new benchmark for Czech ABSA, and our proposed translation–alignment approach offers a scalable solution for adapting ABSA resources to other low-resource languages.

Keywords: Aspect-based sentiment analysis, Large language models, Pre-trained language models, Sentiment analysis, Opinion mining, Czech language

1. Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task that extracts detailed information about entities and aspects. ABSA involves four sentiment elements (Zhang et al., 2023; Šmíd and Kral, 2025): aspect term (a), aspect category (c), sentiment polarity (p), and opinion term (o). For example, in *"Delicious tea"*, these correspond to *"tea"*, *"drinks quality"*, *"positive"*, and *"Delicious"*, respectively. Aspect and opinion terms may also be implicit, commonly annotated as *"NULL"*; for instance, in *"Tasty"*, the aspect term is implicit.

Task	Output	Example output
ASTE	$\{(a, o, p)\}$	{("tea", "delicious", POS)}
ASQP	$\{(a, c, o, p)\}$	{("tea", drinks, "delicious", POS)}
ACOS	$\{(a, c, o, p)\}$	{("tea", drinks, "delicious", POS), ("soup", food, "NULL", NEG)}

Table 1: Output format for selected ABSA tasks that involve opinion terms for the input sentence: *"The tea was delicious, unlike soup"*.

Research has gradually evolved from simple ABSA tasks (e.g. aspect term extraction) to compound tasks requiring linked predictions of multiple sentiment elements. More recently, opinion terms have gained increasing attention, with tasks such as aspect sentiment triplet extraction (ASTE) (Peng et al., 2020), aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a), and aspect category opinion sentiment prediction (ACOS) (Cai et al., 2021). Table 1 provides input and output examples for such tasks. Modern approaches often cast these tasks as text generation problems using pre-trained sequence-to-sequence models (Zhang et al., 2021c; Gou et al., 2023).

Large language models (LLMs) like LLaMA 3.1 (Dubey et al., 2024) have advanced natural language processing substantially. While smaller fine-tuned Transformer-based models still outperform LLMs on ABSA when sufficient training data is available (Gou et al., 2023; Zhang et al., 2024), recent studies highlight the potential of fine-tuned LLMs for ABSA (Šmíd et al., 2024a; Šmíd et al., 2026). Moreover, parameter-efficient methods such as QLoRA (Dettmers et al., 2023) make LLM fine-tuning feasible on limited hardware.

Over time, several datasets have been developed for ABSA, including SemEval-2014–2016 (Pontiki et al., 2014, 2015, 2016) and SentiHood (Saeidi et al., 2016). Most focus on English, with SemEval-2016 also covering several other languages. Later extensions introduced opinion term annotations, enabling tasks such as ASTE (Peng et al., 2020; Xu et al., 2020), ASQP (Zhang et al., 2021a), and ACOS (Cai et al., 2021). For Czech, existing ABSA datasets (Steinberger et al., 2014; Hercig et al., 2016; Tamchyna et al., 2015; Šmíd et al., 2024b) support simple or compound tasks but lack opinion term annotations, restricting research

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7973–7984
11–16 May 2026.

on more complex setups. Since the datasets labelled with all four sentiment elements are only available in English, it is not possible to perform cross-lingual comparisons with other languages.

To address this gap, we introduce a new Czech dataset with opinion term annotations, supporting three compound tasks (ASTE, ASQP, ACOS). To our knowledge, it is the first dataset beyond English to allow quadruplet-level tasks such as ASQP and ACOS. The dataset and code is publicly released to foster further research[1]. We benchmark modern Transformer-based models and LLMs in monolingual, multilingual, and cross-lingual settings. To the best of our knowledge, we are the first to explore cross-lingual configurations for all three tasks. Through error analysis, we further highlight the challenges posed by complex ABSA in Czech.

Our main contributions are as follows: 1) We present a new Czech dataset tailored for compound ABSA tasks, complete with opinion term annotations. 2) We evaluate leading large language models in zero-shot, few-shot, and fine-tuning scenarios, analysing their strengths and limitations. 3) We compare fine-tuned LLMs with a multilingual sequence-to-sequence baseline. 4) We propose a novel method for cross-lingual transfer based on data translation and label alignment with LLMs. 5) We conduct an error analysis to highlight the main challenges posed by the dataset and future research directions.

2. Related Work

We review existing ABSA datasets with opinion term annotations, ABSA datasets for Czech, and prior work on ABSA methods, with an emphasis on Czech.

2.1. ABSA Datasets

The USAGE dataset (Klinger and Cimiano, 2014) provides English and German product reviews annotated with aspect and opinion terms. However, the annotations are not linked, which limits the dataset's suitability for tasks requiring aspect–opinion pairing.

Most English ABSA resources build on the SemEval 2014–2016 datasets (Pontiki et al., 2014, 2015, 2016), which focus on the restaurant domain and include aspect terms but no opinion terms. Fan et al. (2019) added opinion terms and linked them to aspect terms, enabling aspect–opinion pair extraction, though sentiment polarity and aspect category were omitted. Peng et al. (2020) extended this work by merging annotations, reintroducing polarity, and producing data suitable for ASTE. Both works exclude sentences with implicit aspects.

Quadruplet-level datasets further enriched the annotations. The ASQP dataset (Zhang et al., 2021a) reintroduced implicit aspects and added aspect categories. The ACOS dataset (Cai et al., 2021) went further by including implicit opinion terms (ASQP only has explicit ones) and expanding coverage to laptops as well as restaurants.

For Czech ABSA, resources remain scarce. The earliest dataset (Steinberger et al., 2014) provides restaurant reviews in the SemEval-2014 format. Tamchyna et al. (2015) created a dataset of IT product reviews with aspect and sentiment annotations. Hercig et al. (2016) expanded the Czech restaurant dataset, retaining the SemEval-2014 format but without linking aspects to categories. More recently, Šmíd et al. (2024b) introduced a Czech dataset in the SemEval-2016 format with linked aspect terms, categories, and sentiment polarities, enabling compound tasks. However, no Czech ABSA dataset includes opinion term annotations.

2.2. Aspect-Based Sentiment Analysis

Early Czech ABSA work (Steinberger et al., 2014; Tamchyna et al., 2015; Hercig et al., 2016) relied on traditional machine learning methods such as maximum entropy classifiers, later replaced by neural networks (Lenc and Hercig, 2016). More recent studies adopt Transformer-based models (Vaswani et al., 2017), including prompt-based approaches (Šmíd and Přibáň, 2023), multitask learning (Přibáň and Pražák, 2023), and advanced fine-tuned architectures (Šmíd et al., 2024b). This mirrors developments in English ABSA, where recent work often frames the task as text generation using sequence-to-sequence models (Zhang et al., 2021c,a; Gao et al., 2022; Mao et al., 2022; Gou et al., 2023; Xianlong et al., 2023).

Large language models have also been explored. In zero- and few-shot settings, they typically underperform compared to fine-tuned baselines (Gou et al., 2023; Zhang et al., 2024), but fine-tuned LLMs have shown strong results in English (Šmíd et al., 2024a), Czech (Šmíd et al., 2026), and other languages (Šmíd et al., 2025a,b; Wu et al., 2025).

3. Dataset Construction

We build upon the `CsRest-M` dataset (Šmíd et al., 2024b), which contains annotations for aspect terms, aspect categories, and sentiment polarity triplets, and is already divided into training, validation, and test sets. Our primary enhancement is the addition of opinion term annotations, which extend the dataset to support the ASTE, ASQP, and ACOS tasks.

[1]Code and dataset are available at the anonymous repository: https://github.com/biba10/Czech-ABSA-Opinion-Dataset-Benchmark

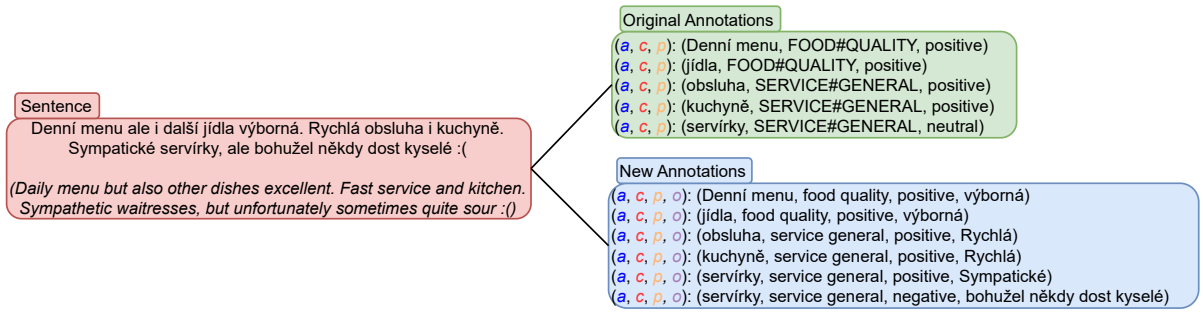


Figure 1: Example of the original annotations (top right) and the updated versions after our modifications (bottom right).

3.1. Annotation Process

Before starting, the annotators developed detailed guidelines, drawing inspiration from the USAGE dataset (Klinger and Cimiano, 2014) as well as the English ACOS (Cai et al., 2021) and ASQP (Zhang et al., 2021a) datasets. The final version of these guidelines is described in the following section, where we summarize the key principles. The development process was iterative: after annotating a few hundred samples, the annotators reviewed their work, discussed any ambiguities, and refined the guidelines accordingly. This approach promoted high inter-annotator agreement and resolved most issues early in the process.

In most cases, the original triplets were retained, with annotators tasked with adding corresponding opinion terms. However, certain sentences required the addition of new quadruplets or adjustments to sentiment polarity. For instance, the phrase *"Velmi rychlá a milá obsluha"* (*"Very fast and friendly service"*) was originally annotated with a single triplet (*"obsluha"*, *"service general"*, *"positive"*) for the aspect term *"obsluha"* (*"service"*). In our dataset, this was expanded to two quadruplets: one linking the opinion term *"Velmi rychlá"* (*"Very fast"*) and another with *"milá"* (*"friendly"*). Similarly, for the phrase *"Sympatické servírky, ale bohužel někdy dost kyselé"* (*"Sympathetic waitresses, but unfortunately sometimes quite sour"*), the original annotation marked the sentiment as *"neutral"* due to its mixed nature for the aspect term *"servírky"* (*"waitresses"*). We refined this by creating two quadruplets: one with *"positive"* polarity for the opinion term *"Sympatické"* (*"Sympathetic"*), and another with *"negative"* polarity for the opinion term *"bohužel někdy dost kyselé"* (*"unfortunately sometimes quite sour"*).

A key decision was whether to include modifiers for opinion terms, such as *"velmi"* (*"very"*) in *"velmi rychlá"* (*"very fast"*). Existing datasets vary: the USAGE dataset (Klinger and Cimiano, 2014) includes modifiers, while ASTE and ACOS datasets (Fan et al., 2019; Peng et al., 2020; Cai et al., 2021) do not. The ASQP dataset (Zhang et al., 2021a) is inconsistent, sometimes including modifiers and sometimes not. We chose to annotate modifiers for two reasons: (1) in Czech, modifiers can significantly affect sentiment intensity, distinguishing mildly positive/negative from strongly positive/negative sentiment, with the former annotated as neutral in related work (Pontiki et al., 2015, 2016; Šmíd et al., 2024b); and (2) modifiers may support future extensions of sentiment polarity annotations, such as introducing *"very positive"* or *"very negative"* categories. This decision may complicate cross-lingual experiments with English ASTE, ASQP, and ACOS datasets, where modifiers are generally omitted. We also annotated implicit opinion terms (marked as *"NULL"*) to make the dataset suitable for the ACOS task.

The main annotation tasks for each review sentence were as follows:

> **Identify opinion terms for each triplet:** Assign an opinion term to each annotated triplet. If no explicit opinion term exists, use *"NULL"*. Introduce additional quadruplets if multiple opinion terms exist for a single aspect term. Adjust sentiment polarity if conflicting opinions are expressed.

After completing the annotation process, we applied two further modifications to ensure consistency with related datasets. First, the aspect categories in the original dataset were provided in the *ENTITY#ATTRIBUTE* format, as in the SemEval datasets. To align with English datasets for ACOS, ASQP, and ASTE, we replaced the "#" symbol with a space and converted the categories to lowercase (e.g. *"FOOD#QUALITY"* became *"food quality"*). Second, we normalized the text to improve readability and processing, including reducing multiple consecutive spaces to a single space and inserting spaces to separate punctuation from words, following conventions used in English datasets. Figure 1 illustrates the transition from the original annotations to the updated versions.

Two native Czech speakers with prior experience in ABSA annotation performed the annotation. A

	Category	Train				Dev				Test				Total			
		Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot
ACOS	ambience general	349	97	30	476	40	12	3	55	121	46	8	175	510	155	41	706
	ambience general	349	97	30	476	40	12	3	55	121	46	8	175	510	155	41	706
	drinks prices	9	14	5	28	1	1	4	6	4	10	0	14	14	25	9	48
	drinks quality	191	50	20	261	24	5	2	31	66	14	4	84	281	69	26	376
	drinks style_options	33	21	7	61	3	2	0	5	19	4	0	23	55	27	7	89
	food prices	41	56	18	115	7	5	2	14	9	21	13	43	57	82	33	172
	food quality	942	400	108	1,450	89	46	7	142	341	117	37	495	1,372	563	152	2,087
	food style_options	137	118	23	278	15	8	6	29	39	53	4	96	191	179	33	403
	location general	30	1	0	31	1	3	0	4	18	3	0	21	49	7	0	56
	restaurant general	544	241	30	814	59	38	5	102	198	101	15	314	801	380	50	1,231
	restaurant miscellaneous	29	63	13	105	3	7	2	12	9	27	7	43	41	97	22	160
	restaurant prices	57	48	27	132	3	3	2	8	25	14	9	48	85	65	38	188
	service general	607	326	53	986	61	46	8	115	217	132	18	367	885	504	79	1,468
	Total	2,969	1,435	334	4,738	306	176	41	523	1,066	542	115	1,723	4,341	2,153	490	6,984
ASQP	ambience general	326	65	19	410	38	12	2	52	114	37	8	159	478	114	29	621
	ambience general	326	65	19	410	38	12	2	52	114	37	8	159	478	114	29	621
	drinks prices	6	8	2	16	1	1	4	6	4	7	0	11	11	16	6	33
	drinks quality	167	33	17	217	23	5	2	30	61	9	3	73	251	47	22	320
	drinks style_options	27	11	4	42	3	2	0	5	16	3	0	19	46	16	4	66
	food prices	33	13	14	60	7	3	2	12	8	11	11	30	48	27	27	102
	food quality	865	260	86	1,211	83	36	7	126	327	85	35	447	1,275	381	128	1,784
	food style_options	106	51	12	169	11	7	3	21	34	31	1	66	151	89	16	256
	location general	24	0	0	24	0	3	0	3	12	2	0	14	36	5	0	41
	restaurant general	439	141	14	594	48	25	2	75	170	57	10	237	657	223	26	906
	restaurant miscellaneous	19	10	4	33	1	2	0	3	5	4	0	9	25	16	4	45
	restaurant prices	49	28	16	93	2	2	2	6	23	8	5	36	74	38	23	135
	service general	558	151	28	737	58	27	5	90	206	72	11	289	822	250	44	1,116
	Total	2,619	771	216	3,606	275	125	29	429	980	326	84	1,390	3,874	1,222	329	5,425
ASTE	Total	2,169	603	178	2,950	234	99	25	358	808	258	64	1,130	3,211	960	267	4,438

Table 2: Detailed statistics of our datasets by aspect category and sentiment polarity. Columns Pos, Neg, Neu, and Tot denote counts for positive, negative, neutral, and total instances, respectively.

third annotator, also experienced in ABSA, assisted in reviewing and resolving disagreements, supporting the other two annotators to ensure consistency and high quality. The final ACOS dataset contains 3,000 sentences with almost 7,000 annotated quadruplets. No instances of content that could be considered discriminatory or overtly racist were found in the dataset, though some reviews contain mild offensive language typical of user-generated reviews.

3.2. Derived ASTE and ASQP Datasets

From the ACOS dataset, we derived task-specific variants. For ASQP, we removed quadruplets with implicit opinion terms (*"NULL"*) and filtered out sentences containing no remaining quadruplets after this exclusion. For ASTE, we omitted aspect category annotations, excluded triplets with implicit aspect or opinion terms, merged identical triplets, and removed sentences without any remaining triplets after these steps.

3.3. Dataset Statistics

Table 2 presents a detailed distribution of our datasets, including sentiment polarities and, for ASQP and ACOS, aspect categories. The most frequent polarity is *"positive"*, while *"neutral"* is relatively rare, accounting for only about 6% of the tuples. The three most common aspect categories are *"food quality"*, *"restaurant general"*, and *"service general"*, whereas the least frequent are *"location general"*, *"drinks prices"*, and *"drinks style_options"*.

Table 3 compares our datasets with existing English restaurant-domain datasets. Our datasets are substantially larger, both in the number of sentences and, more importantly, in the number of annotated tuples. For instance, our ASTE dataset contains nearly twice as many tuples as its English counterpart (4,438 vs 2,247). The distribution of implicit aspect and opinion terms in our ACOS and ASQP datasets is similar to that in English datasets, differing by only a few percentage points.

3.4. Inter-annotator Agreement

Following prior work (Steinberger et al., 2014; Pontiki et al., 2016; Hercig et al., 2016; Cai et al., 2021; Šmíd et al., 2024b), we measure inter-annotator agreement (IAA) using strict quadruplet-matching F1, treating one annotator's labels as gold and the other's as predictions. This metric is standard in ABSA, where annotations consist of structured, multi-label quadruplets with variable cardi-

Dataset	Lang	Split	Sentences	Tuples	IA	IO	IA & IO
ACOS En: (Cai et al., 2021)	Cs (Ours)	Train	2,018	4,738	1,038	1,132	389
		Dev	230	523	104	94	34
		Test	752	1,723	367	333	113
		Total	3,000	6,984	1,509	1,559	536
	En	Train	1,530	2,484	607	448	233
		Dev	171	261	60	54	26
		Test	583	916	213	198	91
		Total	2,283	3,660	880	700	350
ASQP En: (Zhang et al., 2021a)	Cs (Ours)	Train	1,594	3,606	649	0	0
		Dev	198	429	70	0	0
		Test	616	1,390	254	0	0
		Total	2,408	5,425	973	0	0
	En	Train	1,264	1,989	446	0	0
		Dev	316	507	104	0	0
		Test	544	799	179	0	0
		Total	2,124	3,295	729	0	0
ASTE En: (Peng et al., 2020)	Cs (Ours)	Train	1,321	2,950	0	0	0
		Dev	161	358	0	0	0
		Test	505	1,130	0	0	0
		Total	1,987	4,438	0	0	0
	En	Train	857	1,394	0	0	0
		Dev	210	339	0	0	0
		Test	326	514	0	0	0
		Total	1,393	2,247	0	0	0

Table 3: Comparison of our datasets with English counterparts from the SemEval-2016 restaurant domain (Pontiki et al., 2016). IA = implicit aspect terms, IO = implicit opinion terms, IA & IO = tuples containing both an implicit aspect and an implicit opinion term.

nality per instance, making chance-corrected coefficients such as Cohen's kappa less appropriate. After the first 100 examples, IAA was 63%. Following guideline refinement and discussion of annotation issues, it rose to 76% on the subsequent 100 examples. The final IAA over the entire dataset is 85%, indicating substantial agreement and remaining comparable to previously reported results for English datasets (Cai et al., 2021).

Most disagreements stemmed from opinion term annotations, which were the main addition to the original dataset. The primary challenge concerned implicit opinion terms, where annotators sometimes disagreed on whether to treat an expression as explicit (e.g. annotating *"doporučuji"* (*"recommended"*)), or leave it implicit. Other disagreements involved partially overlapping opinion terms, where one annotator included more words than the other, and modifiers that were occasionally omitted during the early stages. Most of these, along with additional early-phase inconsistencies, such as failing to create separate quadruplets for multiple opinion terms, were mitigated through iterative refinement of the guidelines (see Section 3.1 for details).

4. Experiments & Setup

We evaluate our new Czech ABSA dataset on three tasks: ASTE, ACOS, and ASQP. The primary evaluation metric is micro F1-score. For fine-tuning experiments, results are averaged over five runs with different random seeds, whereas zero-shot and few-shot experiments are performed with a single run, as they do not involve stochastic parameter updates. A predicted tuple is considered correct only if all of its components exactly match the corresponding gold tuple.

4.1. Sequence-to-Sequence Models

Following prior work on Czech ABSA (Šmíd and Přibáň, 2023; Šmíd et al., 2024b), we fine-tune large mT5 (Xue et al., 2021), Transformer-based encoder–decoder (sequence-to-sequence) model. The encoder produces contextualized representations $\mathbf{e}$ of the input text. The decoder estimates $P_{\Theta}(y|\mathbf{e})$ over the output sequence y, generating each token y_i conditioned on $\mathbf{e}$ and the previously generated tokens before the i-th step, denoted as $y_{<i}$.

We transform ABSA labels into a textual format using six special tokens: `<aspect>` for aspect terms, `<opinion>` for opinion terms, `<category>` for aspect categories, `<polarity>` for sentiment polarities, `<null>` for implicit terms, and `<ssep>` to separate tuples. We abbreviate sentiment polarity to its first three letters (e.g. *"pos"* for *"positive"*), and order the sentiment elements as $a \rightarrow o \rightarrow c \rightarrow p$ (Gou et al., 2023). Figure 2 illustrates the conversion process for a given input. The ASTE task does not require the `<null>` and `<category>` tokens.

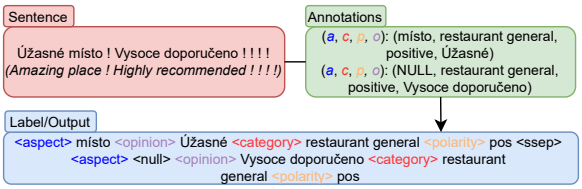


Figure 2: Example of converting ABSA annotations into output sequences for sequence-to-sequence models. Special tokens represent aspect terms, opinion terms, categories, sentiment polarities, and tuple separators.

During fine-tuning, we update all model parameters and minimize the negative log-likelihood as

$$\mathcal{L} = -\sum_{i=1}^{n} \log p_{\Theta}(y_i|\mathbf{e}, y_{<i}), \quad (1)$$

where n is the length of the target sequence y.

4.2. Large Language Models

We evaluate decoder-only LLMs in zero-shot, few-shot, and fine-tuning settings. Unlike encoder–decoder models, decoder-only models generate tokens autoregressively without a dedicated

encoder. Their training loss follows the same principle as sequence-to-sequence models, but conditions only on previously generated tokens.

We adopt prompts from prior Czech ABSA work (Šmíd et al., 2024a; Šmíd et al., 2026). Few-shot prompts use the first ten training examples, which provide sufficient coverage for fair comparison (Šmíd et al., 2026). These examples are representative and unsorted, ensuring diversity beyond sentiment cues. Figure 3 shows an ACOS prompt with one demonstration and its expected output. We adapt the same format for ASQP (without implicit opinions) and ASTE (excluding aspect categories and implicit terms). Prompts are written in English, as prompt language has little impact on Czech ABSA performance (Šmíd et al., 2026).

According to the following sentiment elements definition:
- The "aspect term" refers to a specific feature, attribute, or aspect of a product or service on which a user can express an opinion. Explicit aspect terms appear explicitly as a substring of the given text. The aspect term might be "null" for the implicit aspect.
- The "aspect category" refers to the category that aspect belongs to, and the available categories include: "ambience general", "drinks prices", "drinks quality", "drinks style_options", "food prices", "food quality", "food style_options", "location general", "restaurant general", "restaurant miscellaneous", "restaurant prices", "service general".
- The "sentiment polarity" refers to the degree of positivity, negativity or neutrality expressed in the opinion towards a particular aspect or feature of a product or service, and the available polarities include: "positive", "negative" and "neutral". "neutral" means mildly positive or mildly negative. Quadruplets with objective sentiment polarity should be ignored.
- The "opinion term" refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service. Explicit opinion terms appear explicitly as a substring of the given text. The opinion term might be "null" for the implicit opinion.

Please carefully follow the instructions. Ensure that aspect terms are recognized as exact matches in the review or are "null" for implicit aspects. Ensure that aspect categories are from the available categories. Ensure that sentiment polarities are from the available polarities. Ensure that opinion terms are recognized as exact matches in the review or are "null" for implicit opinions.

Recognize all sentiment elements with their corresponding aspect terms, aspect categories, sentiment polarity, and opinion terms in the given input text (review). Provide your response in the format of a Python list of tuples: 'Sentiment elements: [("aspect term", "aspect category", "sentiment polarity", "opinion term"), ...]'. Note that ", ..." indicates that there might be more tuples in the list if applicable and must not occur in the answer. Ensure there is no additional text in the response.

Input: """Úžasné místo ! Výsoce doporučeno ! ! ! !"""
Sentiment elements: [("místo", "restaurant general", "positive", "Úžasné"), ("null", "restaurant general", "positive", "Vysoce doporučeno")]

Input: """Vepřové koleno bylo skvělé ."""
Sentiment elements: [("Vepřové koleno", "food quality", "positive", "skvělé")]

Figure 3: Prompt for the ACOS task with example input, expected output (green box), and one demonstration (dashed box, used only in few-shot scenarios).

We fine-tune LLMs with QLoRA (Dettmers et al., 2023), which applies 4-bit quantization to a frozen backbone and learns only a small set of LoRA weights (Hu et al., 2022), substantially reducing memory requirements.

We evaluate a range of models for zero-shot and few-shot experiments: GPT-4o mini (OpenAI, 2024), Orca 2 (13B) (Mitra et al., 2023), LLaMA 3.1 (8B, 70B), LLaMA 3.3 (70B) (Dubey et al., 2024), Gemma 3 (4B, 12B, 27B) (Team et al., 2025), and Aya 23 (8B, 35B) (Aryabumi et al., 2024). Fine-tuning is restricted to models up to 12B parameters and uses the prompt shown in Figure 3. All models are open-source except GPT-4o mini.

4.3. Multilingual and Cross-lingual Experiments

Beyond monolingual setups, we perform multilingual and cross-lingual experiments. For English, we use datasets derived from the SemEval-2016 restaurant domain (Pontiki et al., 2016): ASTE (Peng et al., 2020), ACOS (Cai et al., 2021), and ASQP (Zhang et al., 2021a).

In multilingual experiments, models are trained on the union of Czech and English data and evaluated on Czech test sets, with model selection based on combined validation performance across both languages.

In cross-lingual settings, English serves as the source language and Czech as the target. Model selection is based on English validation data. We test two variants: (1) fine-tuning only on the English dataset, and (2) fine-tuning on English plus its machine-translated Czech counterpart.

To create aligned translations, we use GPT-4o mini with detailed instructions (Figure 4). Because translation alters word counts and positions, we filter outputs by tuple counts, categories, sentiment polarities, and implicit/explicit markers. The primary source of errors – explicit terms not matching the translated text – occurred in about 10% of cases. To our knowledge, this is the first use of LLMs for ABSA data translation with alignment. Compared to traditional methods such as FastAlign (Dyer et al., 2013), which is often unreliable, or symbol-marking (Zhang et al., 2021b), which can drop or misalign labels, our approach provides a more robust alternative.

Your task is to translate the given review for ABSA from English to Czech and adjust the labels accordingly.
The review is in the format: <text>####<label>
Label is a list of lists, each list consists of 4 elements: aspect term, aspect category, sentiment polarity, opinion term.

Instructions:
1. Translate the review text to Czech.
2. Adjust the labels so that:
  - The number of labels remains the same.
  - Aspect categories and sentiment polarities are copied directly from the original English labels (do not translate these).
  - Aspect terms and opinion terms are extracted from the translated review in Czech (based on their actual occurrences in the translated text, not by translating them separately).
  - For `NULL` aspect or opinion terms, leave them unchanged in the label.
3. Return the result in the following format: `<translated_review>####<adjusted_labels>`.
  - Ensure the translated review and adjusted labels are separated by `####`.
  - Labels must be a list of lists, each containing four elements in this order: `[aspect term, aspect category, sentiment polarity, opinion term]`.

Examples:
Input: The food was great####[['food', 'food quality', 'positive', 'great']]
Output: Jídlo bylo skvělé####[['jídlo', 'food quality', 'positive', 'skvělé']]
Ensure proper grammar and naturalness in the Czech translation.
Return only the `<translated_review>####<adjusted_labels>` format as output, without any additional comments or text.

Figure 4: Prompt for translating the ACOS dataset from English to Czech with aligned labels. The full prompt contains five different representative input/output examples.

4.4. Hyperparameters

We employ all open-source models from the HuggingFace Transformers library[2] (Wolf et al., 2020). All experiments use greedy search decoding and run on a single NVIDIA L40 GPU with 48 GB memory. Considering preprocessing, we lowercase all sentences and labels to maintain consistency across monolingual, multilingual, and cross-lingual experiments, preventing mismatches between datasets (some already lowercased) and mitigates performance degradation caused by case differences.

We fine-tune mT5 for 20 epochs with batch size 16, learning rate 1e-4, and the AdamW optimizer (Loshchilov and Hutter, 2019), selected for consistent validation performance across tasks.

For QLoRA fine-tuning, we use 4-bit NormalFloat (NF4) with double quantization and bf16 computation. Training parameters include batch size 16, constant learning rate $2e-4$, AdamW optimizer, and LoRA applied to all Transformer linear layers, with $r = 64$ and $\alpha = 16$. For Gemma 3, we follow prior recommendations (Šmíd et al., 2026) and use $r = 64$ and $\alpha = 128$. Training runs for up to 5 epochs, selecting the best model by validation loss. Loss is computed only on model-generated tokens, excluding prompts (Mitra et al., 2023). Zero-shot and few-shot models are quantized to 4 bits, as they perform similarly to full-precision models (Dettmers et al., 2023; Šmíd et al., 2024a).

4.5. Results

This section presents the results. Among the three tasks, ASTE consistently achieves the highest scores, followed by ASQP, with ACOS being the most challenging. This pattern reflects the increasing complexity of extracting fine-grained sentiment information.

Table 4 shows the monolingual results. GPT-4o mini achieves the strongest zero-shot results, followed by Gemma 3 27B and LLaMA 3.3 70B, with the latter substantially outperforming its predecessor LLaMA 3.1 70B. Larger and more recent models consistently outperform smaller or older ones. Few-shot examples generally improve performance, though GPT-4o mini shows a slight decrease. Gemma 3 27B achieves the best few-shot results, while Orca 2 13B benefits most from demonstrations, surpassing Aya 23 8B and LLaMA 3.1 8B despite being English-centric. Aya 23 8B consistently outperforms LLaMA 3.1 8B, likely due to its official multilingual support, which includes Czech. Fine-tuned models outperform all zero- and few-shot LLMs, with mT5 achieving the best overall results. Fine-tuned mT5 surpasses the strongest few-shot LLMs by 15–25% across tasks while offering efficiency advantages in memory and inference speed.

Type	Model	ASTE	ASQP	ACOS	AVG
Zero-shot	Aya 23 8B	25.59	8.76	5.97	13.44
	Aya 23 35B	37.34	26.00	23.18	28.84
	Gemma 3 4B	30.87	15.20	8.23	18.10
	Gemma 3 12B	51.64	31.75	28.00	37.13
	Gemma 3 27B	53.27	41.27	**35.43**	43.33
	LLaMA 3.1 8B	29.89	7.37	3.50	13.59
	LLaMA 3.1 70B	45.94	31.86	26.49	34.76
	LLaMA 3.3 70B	50.21	38.29	32.08	40.19
	Orca 2 13B	15.24	10.21	8.09	11.18
	GPT-4o mini	**56.43**	**42.12**	33.77	**44.10**
Few-shot	Aya 23 8B	35.73	28.37	20.65	28.25
	Aya 23 35B	42.79	37.10	30.67	36.85
	Gemma 3 4B	46.72	33.37	25.07	35.05
	Gemma 3 12B	54.49	40.17	33.41	42.69
	Gemma 3 27B	53.27	**46.91**	**39.35**	**46.51**
	LLaMA 3.1 8B	29.80	17.02	14.20	20.34
	LLaMA 3.1 70B	48.23	40.99	33.37	40.86
	LLaMA 3.3 70B	**55.02**	44.36	33.76	44.38
	Orca 2 13B	40.56	28.83	21.41	30.27
	GPT-4o mini	54.76	40.68	33.05	42.83
Fine-tuning	mT5	**70.74**	**64.09**	**58.06**	**64.30**
	Aya 23 8B	69.87	61.85	57.46	63.06
	Gemma 3 4B	58.18	48.90	44.54	50.54
	Gemma 3 12B	65.23	58.56	48.45	57.41
	LLaMA 3.1 8B	66.70	60.25	56.18	61.04
	Orca 2 13B	67.13	57.93	53.85	59.64

Table 4: F1 scores in monolingual settings for different models and tasks, alongside with an average result across tasks. **Bold** results indicate the best result for each combination of "Type" and task.

Type	Model	ASTE	ASQP	ACOS	AVG
Original	mT5	49.60	41.68	35.17	42.15
	Aya 23 8B	51.02	42.12	36.41	43.18
	Gemma 3 4B	49.08	36.70	30.80	38.86
	Gemma 3 12B	**53.40**	**44.17**	37.52	**45.03**
	LLaMA 3.1 8B	47.29	43.35	36.56	42.40
	Orca 2 13B	48.09	41.90	**38.32**	42.77
Translated	mT5	51.00	40.07	35.85	42.31
	Aya 23 8B	51.59	40.56	40.46	44.20
	Gemma 3 4B	48.63	39.32	32.82	40.26
	Gemma 3 12B	**57.98**	**49.13**	**41.11**	**49.41**
	LLaMA 3.1 8B	49.20	44.68	37.91	43.93
	Orca 2 13B	49.29	36.62	38.68	41.53

Table 5: F1 scores in cross-lingual settings for different models and tasks, alongside with an average result across tasks. **Bold** results indicate the best result for each combination of "Type" and task.

Table 5 shows the cross-lingual results. Training on English only ("Original") yields similar results across models, except for Gemma 3 4B, which lags behind. Gemma 3 12B achieves the

[2] https://github.com/huggingface/transformers

highest average scores, while Orca 2 13B performs best on ACOS. Adding translated English-to-Czech data ("Translated") improves several models: Aya 23 8B and LLaMA 3.1 8B gain over 1%, while Gemma 3 improves by over 2% (4B) and 4% (12B). Gemma 3 12B achieves the best overall cross-lingual results. Nevertheless, performance remains 15–20% lower than fine-tuned monolingual results, underscoring the challenges of language transfer and annotation inconsistencies. A significant factor is the treatment of opinion terms: English datasets exclude modifiers, whereas Czech datasets include them, resulting in models trained on English data missing modifiers. Cross-lingual performance, however, often surpasses zero- and few-shot monolingual baselines – for example, Orca 2 13B improves by 30% over its zero-shot and 12% over its few-shot performance.

Model	ASTE	ASQP	ACOS	AVG
mT5	**70.93**	**64.64**	**59.12**	**64.90**
Aya 23 8B	69.28	63.71	56.15	63.05
Gemma 3 4B	66.01	51.44	53.90	57.11
Gemma 3 12B	68.16	61.61	56.74	62.17
LLaMA 3.1 8B	66.70	63.60	57.37	62.56
Orca 2 13B	67.35	57.67	53.20	59.41

Table 6: F1 scores in multilingual settings for different models and tasks, alongside with an average result across tasks. **Bold** results indicate the best result for each task.

Finally, Table 6 presents results of multilingual experiments, where fine-tuned mT5 again achieves the best overall results, followed by Aya 23 8B, LLaMA 3.1 8B, and Gemma 3 13B, with Gemma 3 4B performing worst. Multilingual fine-tuning generally outperforms cross-lingual setups, as models benefit from training data in both languages, but brings little advantage over monolingual fine-tuning.

Overall, the results highlight trade-offs between fine-tuned models and LLMs. Fine-tuning consistently delivers the best performance, with mT5 combining accuracy with memory and inference efficiency. LLMs remain attractive for rapid deployment, as few-shot prompting and multilingual pretraining boost their performance; however, results vary substantially across models. Larger, newer, and multilingual-aware LLMs perform best. Cross-lingual transfer benefits modestly from machine translation, though dataset inconsistencies limit gains, and multilingual setups do not always surpass monolingual ones. Model choice should therefore balance task complexity, data availability, and computational resources, with fine-tuning as the most reliable strategy and LLMs offering flexible multilingual alternatives.

4.6. Error Analysis

To better understand model challenges on our datasets, we manually analyzed 100 randomly sampled examples per dataset for GPT-4o mini (zero-shot), LLaMA 3.3 70B (few-shot), and fine-tuned mT5. Figure 5 summarizes the error distribution.

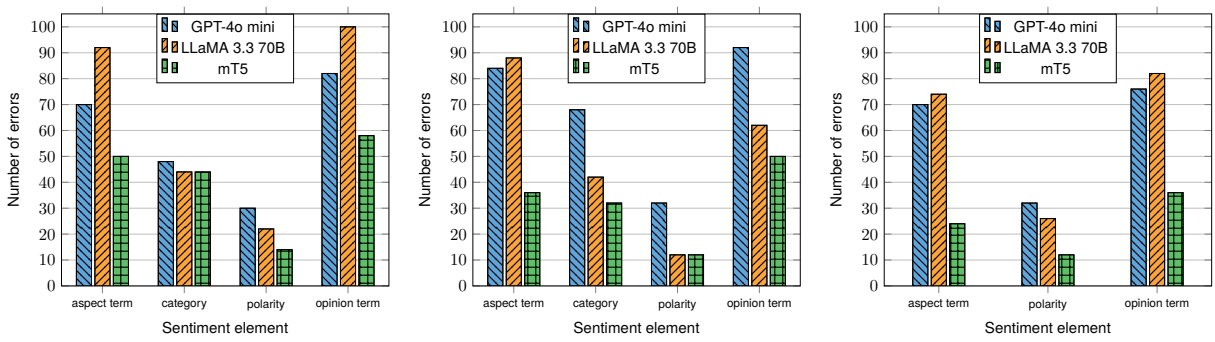
The analysis reveals that opinion terms are the most challenging to predict, followed by aspect terms, while sentiment polarity is the easiest to predict. These conclusions align with findings on English datasets (Zhang et al., 2021a; Šmíd et al., 2024a). One reason is that opinion terms often span multiple words, whereas aspect terms are usually single words. Implicit opinion terms are particularly difficult, explaining why ACOS yields more errors than ASQP or ASTE. Additional challenges arise from typos, abbreviations, and slang in aspect and opinion terms. In contrast, aspect categories and sentiment polarities are drawn from fixed label sets, making them comparatively easier to predict.

Fine-tuned mT5 consistently produced fewer errors, especially for aspect and opinion terms. Some predictions by GPT-4o mini and LLaMA 3.3 70B, however, could be considered acceptable under alternative annotation interpretations. For instance, models occasionally generalized multiple foods into a single aspect term (e.g. *"jídlo"* (*"food"*)), predicted canonical morphological forms (e.g. *"obsluha"* (*"service"*) instead of *"obsluhou"*), included plausible but unannotated aspects such as *price*, or merged multiple opinion terms into one (e.g. *"milá, příjemná"* (*"nice, pleasant"*)) that annotations separate into multiple tuples. Fine-tuning helps align predictions to dataset-specific conventions.

LLMs also produce genuine errors. GPT-4o mini frequently fails to predict any sentiment tuples, while LLaMA 3.3 70B sometimes misclassifies clearly *"positive"* sentiments, such as *"hezké prostředí"* (*"nice environment"*), as *"negative"*. Both models struggle with idiomatic expressions. For example, *"Pivečko jak křen"* (idiomatically expressing positive sentiment toward beer) was correctly segmented into aspect and opinion terms. However, LLaMA 3.3 70B predicted *"negative"* sentiment and GPT-4o mini predicted *"neutral"*, whereas mT5 correctly predicted *"positive"*, likely due to exposure during fine-tuning.

Errors in aspect categories typically involve confusion between semantically related classes, such as *"restaurant general"* vs. *"restaurant miscellaneous"* or *"drinks prices"* vs. *"restaurant prices"*. Rare categories such as *"location general"* are also problematic. For sentiment polarity, the main issue is misclassifying *"neutral"* cases – often mildly positive or negative – as either *"positive"* or *"negative"*, likely due to their low frequency in the dataset.

Generation errors were absent for fine-tuned mT5, contrasting earlier reports (Zhang et al.,



(a) ACOS (b) ASQP (c) ASTE

Figure 5: Number of error types for each dataset for GPT-4o mini in zero-shot settings, LLaMA 3.3 70B in few-shot settings, and fine-tuned mT5.

2021a). LLaMA 3.3 70B and GPT-4o mini produced only 1–2 formatting errors each, indicating generally reliable sequence generation.

5. Conclusion

This paper introduces a manually annotated Czech restaurant dataset for aspect-based sentiment analysis, enriched with opinion term annotations and designed to support three ABSA tasks of varying granularity. Through monolingual, cross-lingual, and multilingual experiments, we evaluate both fine-tuned Transformer-based models and large language models, highlighting strengths and limitations in low-resource settings. Our findings show that fine-tuned models remain the most reliable choice, while LLMs provide flexible alternatives for rapid adaptation across languages.

To address the challenges of cross-lingual transfer, we propose a translation and label alignment methodology using LLMs, which improves performance and offers a scalable alternative to manual annotation. Error analysis further reveals persistent difficulties, such as detecting subtle opinion terms, disambiguating fine-grained aspect categories, and handling nuanced sentiment expressions in the Czech language. These insights point to clear directions for future research on more robust modelling of opinion expressions and multilingual ABSA.

Limitations

While our newly created Czech ABSA dataset provides high-quality annotations, it is limited to the restaurant domain, which may restrict the generalizability of models trained on it to other domains such as hotels, e-commerce, or healthcare. The dataset also focuses exclusively on Czech, so findings may not directly transfer to other languages or dialects. Additionally, implicit opinion terms remain challenging to annotate, and despite review by multiple experienced annotators, some subjectivity may persist, which could affect model performance on subtle sentiment expressions. Finally, the effectiveness of the cross-lingual approach with translation depends on the LLM used for both translation and label alignment.

Ethics Statement

During the creation of the new Czech ABSA dataset, the annotators did not identify any instances of content that could be considered discriminatory or overtly racist. Some reviews contain mild offensive language typical of user-generated text, but no systematic bias or harmful content was present in the dataset itself.

We also note that the models used in this study are pre-trained on large internet corpora. As such, they may exhibit unintended biases related to race, gender, or other sensitive attributes due to the nature of the pre-training data.

Acknowledgements

The work of Jakub Šmíd has been supported by the Grant No. SGS-2025-022 – New Data Processing Methods in Current Areas of Computer Science. The work of Pavel Král has been supported by the project R&D of Technologies for Advanced Digitalization in the Pilsen Metropolitan Area (DigiTech) No. CZ.02.01.01/00/23_021/0008436. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

6. Bibliographical References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos,

Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, Michal Konkol, and Josef Steinberger. 2016. Unsupervised methods to improve aspect-based sentiment analysis in czech. *Computación y Sistemas*, 20(3):365–375.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Roman Klinger and Philipp Cimiano. 2014. The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ladislav Lenc and Tomás Hercig. 2016. Neural networks for sentiment analysis in czech. In *Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016)*, volume 1649 of *CEUR Workshop Proceedings*, pages 48–55, Bratislava, Slovakia. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, CreateSpace Independent Publishing Platform.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

OpenAI. 2024. *GPT-4o*. Accessed November 2024.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Pavel Přibáň and Ondřej Pražák. 2023. Improving aspect-based sentiment with end-to-end semantic role labeling model. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 888–897, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Jakub Šmíd and Pavel Kral. 2025. Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. *Information Fusion*, 120:103073.

Jakub Šmíd and Pavel Přibáň. 2023. Prompt-based approach for Czech sentiment analysis. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1110–1120, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024a. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.

Jakub Šmíd, Pavel Přibáň, and Pavel Kral. 2025a. Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 757–766. INSTICC, SciTePress.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2025b. LACA: Improving cross-lingual aspect-based sentiment analysis with LLM data augmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–853, Vienna, Austria. Association for Computational Linguistics.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2026. Large language models for czech aspect-based sentiment analysis. In *Text, Speech, and Dialogue*, pages 15–26, Cham. Springer Nature Switzerland.

Jakub Šmíd, Pavel Přibáň, Ondrej Prazak, and Pavel Kral. 2024b. Czech dataset for complex aspect-based sentiment analysis tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4299–4310, Torino, Italia. ELRA and ICCL.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland. Association for Computational Linguistics.

Ales Tamchyna, Ondrej Fiala, and Katerina Veselovská. 2015. Czech aspect-based sentiment analysis: A new dataset and preliminary results. In *ITAT*, pages 95–99.

Gemma Team, Aishwarya Kamath, et al. 2025. Gemma 3 technical report.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, page 6000–6010. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chengyan Wu, Bolei Ma, Zheyu Zhang, Ningyuan Deng, Yanqing He, and Yun Xue. 2025. Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *International Journal of Machine Learning and Cybernetics*.

Luo Xianlong, Meng Yang, and Yihao Wang. 2023. Tagging-assisted generation model with encoder and decoder supervision for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Singapore. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges . *IEEE Transactions on Knowledge & Data Engineering*, 35(11):11019–11038.